REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
01-02-2007	Progress Report	11/1/06 -1/31/07

4. TITLE AND SUBTITLE

Next-Generation Image and Sound Processing Strategies: Exploiting the Biological Model`

5a. CONTRACT NUMBER

N00014-06-1-0746

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

PI: Mel, Bartlett W.

Co-PI's: Grzywacz, Norberto M., Itti, Laurent, Narayanan, Shri

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

University of Southern California
Dept. of Biomedical Engineering
1042 Downey Way, DRB 140
Los Angeles, CA 90089-1111

8. PERFORMING ORGANIZATION REPORT NUMBER

95-164-2394

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Office of Naval Research
San Diego Regional Office
140 Sylvester Road
San Diego, CA 92106-3521

10. SPONSOR/MONITOR'S ACRONYM(S)

ONR

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

N00014-06-1-0746

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

N/A

14. ABSTRACT

14. The main objective of this project is to extend the technical state-of-the-art in mid-level visual and auditory signal processing using an integrative biologically inspired approach. Though our research and development efforts are focused on different levels of sensory information processing, from low-level sensory adaptation to object selection and recognition, all of our efforts described in this progress report intersect at the level of features: what low level sensory features to extract, what methods used to extract them, and how to adapt feature detector parameters, such as their gains, in order to perform optimally in changing environments. In this progress report, we detail developments in each of these areas.

15. SUBJECT TERMS

15. Mid-level visual and auditory processing, biologically inspired, feature extraction, sensory adaptation, attentional focus

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Bartlett W. Mel
none	none	none	none	8	19b. TELEPHONE NUMBER *(Include area code)* 213-740-0334

Progress Report for ONR grant #N00014-06-1-0746

Coverage Period: November 1, 2007 to Jan. 31, 2007

Title: Next-Generation Image and Sound Processing Strategies: Exploiting the Biological Model

PI: Bartlett W. Mel, University of Southern California
Co-PI's: Norberto M. Grzywacz, Laurent Itti, Shri Narayanan

Abstract

The main objective of this project is to extend the technical state-of-the-art in mid-level visual and auditory signal processing using an integrative biologically inspired approach. Though our research and development efforts are focused on different levels of sensory information processing, from low-level sensory adaptation to object selection and recognition, all of our efforts described in this progress report intersect at the level of features: what low level sensory features to extract, what methods used to extract them, and how to adapt feature detector parameters, such as their gains, in order to perform optimally in changing environments. In this progress report, we detail developments in 4 areas:

1. Use of Kalman filters to model as-yet unexplained features of visual speed adaptation in human observers. These experiments bear on optimal adaptation/gain control mechanisms for the extraction of a broad class of low-level features in non-stationary natural environments;

2. Continuing development and testing of low-level image features needed for junction detection, including center-surround, edge, endstop, and corner detectors that include gain control mechanisms (see item 1 above). Junctions (L vs. T vs. Y, etc.) are known to be critical for both 2-D and 3-D object classification, one of the key long-term goals of this project;

3. Refinement of low level features needed for sound classification in both clean and noisy environments; the features we are exploring in the auditory domain are closely related to those used in vision (see items 1 and 2 above), including center-surround and oriented operators; we continue to explore commonalities between the two sensory domains;

4. Continuing development of a vertically integrated system that combines low-level feature extraction, attentional mechanisms, and simple object recognition to control a robot arm engaged in a task.

Recent progress in each of these areas is outlined in the following sections.

1. Speed Adaptation as Kalman Filtering

One of the main weaknesses of artificial scene-understanding systems is their inability to adapt to new environments. If, for example, a robotic submarine moves from shallow waters to deeper waters, the statistics of the visual environment change significantly. The intensity and chromaticity of light changes with depth, and absorption and scattering in water cause the background visual texture to lose high spatial frequencies (Balboa & Grzywacz, 2003). If an imagine analyzer has limited dynamic range, then this device may not be simultaneously optimal in deep and shallow water. Similar difficulties arise on land when dealing with changing lighting and contrast conditions, different backgrounds (desert vs. forest), and so on. The retina and the brain, however, have numerous adaptation mechanisms that allow them to perform near-optimally under changing conditions. The key is to continuously optimize the system's parameters by comparing samples from the environment with predictions based on internal models of the environment. If the samples agree with the predictions, then the model parameters remain constant. If they disagree, parameters are modified according to an optimal protocol prescribed by a generalized Kalman-filtering strategy. This strategy is not limited to speed (Fig. 1) and contrast adaptation (Grzywacz & De Juan, 2003), which we have so far experimentally explored. The approach can be applied as well to other types of visual feature extraction (see item 2), or to systems engaged in sound interpretation (see item 3), as long as good measurements—and good models—of the natural signal statistics can be obtained.

Given its potential importance in understanding sensory adaptation in general, we have been investigating whether the human visual system uses Kalman filtering for sensory adaptation, and in particular, for speed adaptation. Speed adaptation occurs when an observer is exposed to high (or low) optic flow speeds for extended periods of time (e.g. drivers, pilots), and results in an altered perception of flow speeds and in particular a reduction (or increase) in perceived speed under default conditions. To quantify and model this, we performed a speed-matching experiment to evaluate the time course of adaptation to an abrupt velocity change. Experimental results are in good agreement with Kalman-filter predictions for speed adaptation. When subjects adapt to a low speed display that suddenly increases in speed, the time course of adaptation unfolds in two phases with different time constants: a rapid decrease of perceived speed followed by a slower phase (Fig. 1A). In contrast, when speed changes from fast to slow, speed adaptation follows a single time course (Fig. 1B). A Kalman filter model predicts this asymmetry: Low speeds are much more common than high speeds in natural environments, so that a transition to higher speeds, which are rare, provides strong evidence for a shift in environment. This triggers a rapid initial phase of adaptation (followed by a more conservative period of gradual adjustment). In contrast, a sudden transition to lower speeds, which are more common in any given environment, is more consistent with an unchanged environment, leading to a gradual (single time constant) adaptation. Interestingly, we found both in simulations and in psychophysical experiments on human subjects that the difference between slow-to-fast and fast-to-slow adaptation disappears when the adapting stimulus is noisy (Figs. 1C and D).

Based on these validations of the Kalman-filtering model in human subjects, we are exploring ways to implement similar optimal adaptation strategies to improve the performance of our low-level feature extraction methods.

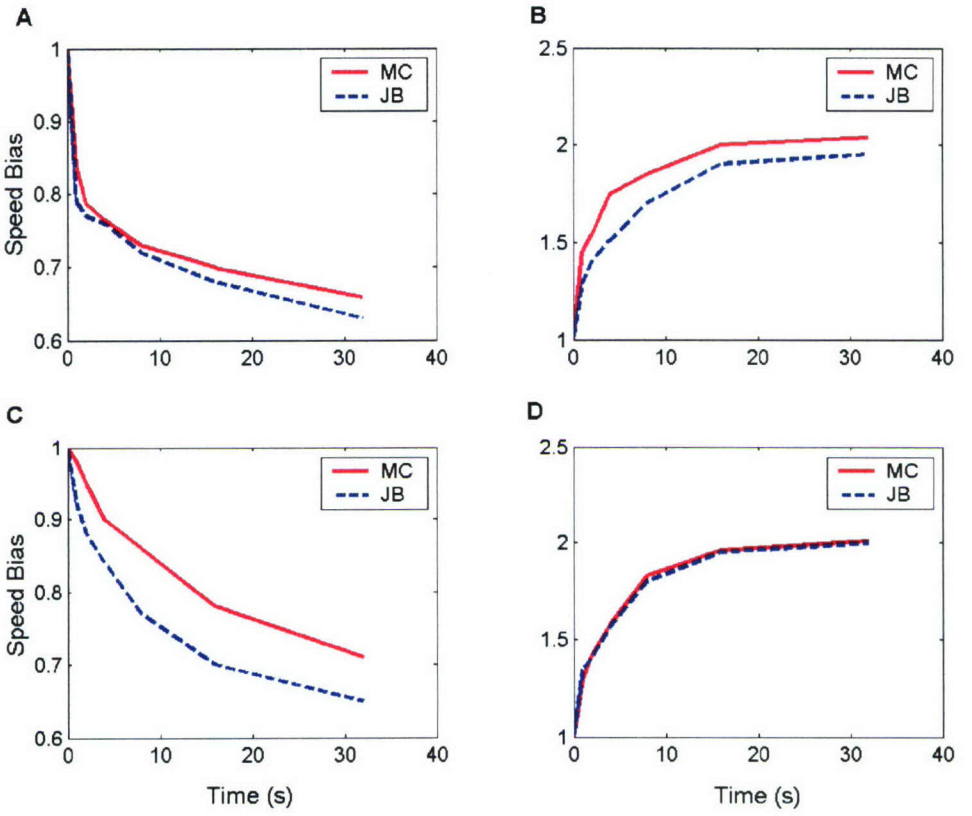

Figure 1. Results of the psychophysical experiments. Each panel shows the speed bias as a function of time for one experimental situation and two subjects. Panels A and B show the results for the noise-free situation. Panel A shows that for slow-to-fast transition, the bias presents two temporal phases, such as those found with model simulations. On the other hand, results show that the bias occurs in a single phase for the fast-to-slow transition (Panel B). As predicted by the model, when we apply noise to the stimulus, the bias occurs in one phase for both transitions (Panels C and D).

2. Feature Extraction for Junction Classification

As discussed in our last progress report, we are developing a hierarchical, nested self-organizing feature map (SOFM) architecture to learn to classify junctions (and eventually objects) in complex visual scenes. Correct classifications of junctions are important given that they are known to be vital for shape-based object recognition. However, reliably extracting junctions in video images has proven to be enormously technically challenging.

As a prelude to junction detection, we have followed the example of biological visual systems by focusing our recent efforts on the development of robust edge, endstop and corner detectors—the ingredients from which junction detectors can be built—and that function well under wide ranging scene conditions. We have recently developed an endstop detector designed to respond strongly to edges that either terminate or abruptly change direction, and a complementary detector with a center-surround organization that responds well to corners (Fig. 2).

Figure 2. Endstopped edge (red) and corner detector (green) applied to an image; lower frame is enlarged to show detail.

The detectors are nonlinear in their construction. For example, the edge and endstop detectors, unlike linear filters, each rely on multiple tests of proper edge structure along the filter's length, each of which is run through a saturating nonlinearity to introduce a degree of image contrast invariance, while the center-surround filter performs a comparable series of radially-oriented tests. Both types of filters also include a divisive normalization operation that implements a form of adaptive gain control similar in spirit to that discussed in Section 1 above. As shown in Fig. 2 (conventional edge responses were omitted for clarity), the endstop and center-surround filters concentrate their activity mostly on appropriate small image structures, with occasional errors that we are investigating and attempting to eliminate. We have noted in previous experiments that junction detectors, because they rely on information from multiple edge and endstop detectors, often respond more cleanly than the pattern of lower-level detector activations might suggest. We expect the same will hold true here as we incorporate these low-level features into our SOFM learning model.

3. Biologically Inspired Speech and Audio Processing

In our previous work, we worked with a detailed early auditory model that mimics the processing stages starting from outer ear to cochlear nucleus via band-pass basilar membrane filters, ear hair cell stages and lateral inhibitory network followed by a leaky integration. It was shown that the derived MFCC-equivalent auditory based features (ABF) outperformed the MFCC features in a speaker independent noisy digit recognition task using Aurora2 database. Also, we used principal component analysis (PCA) to find the important components of the output of leaky-integrate-and-fire (LIF) neuron aiming for noise and dimension reduction. This new feature set (PC-ABF) was beneficial in the recognition task for speech with low SNR, but performed poorly compared to ABF and MFCC for cleaner speech (Fig 1).

Auditory nerves have limited dynamic range. The dynamic range of the basilar membrane and the neural response are compressed nonlinearly by outer hair cells. The outer hair cells provide greater amplification to low signal levels. We modified the early auditory model, and used logarithmic amplitude transformation to model the nonlinear compression due to outer hair cells, and then applied PCA. This new features are called LPCA_ABF. We achieved the best ASR performance by keeping 25 principal components with LPCA_ABF features. The ASR experiment results are summarized in Fig 1. With this new set of features, our ASR performance improved even more for speech with low SNR, and the ASR performance degradation for speech with moderate to high SNR faced with using PC_ABF feature set was also removed. The data used in these experiments contained subway noise. Also, we got similar results in the ASR experiments for speech with babble, car, and exhibition noise given in the Aurora2 database. Our ASR results indicate that the LPCA_ABF provides on average 40% WER improvement over MFCC in noisy speech recognition (averaged over all four types of noise and for all SNR levels).

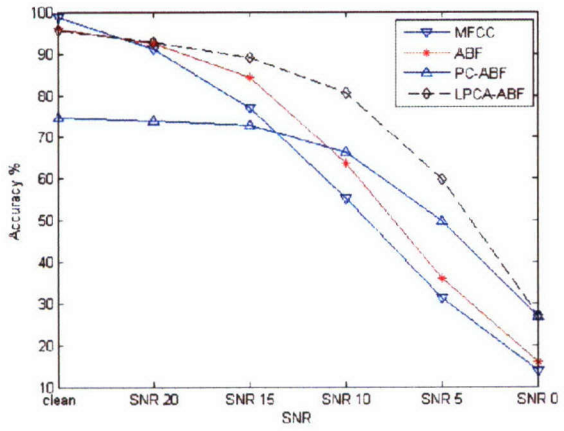

Figure 3. ASR experiment results for speech for with subway noise in a digit recognition task.

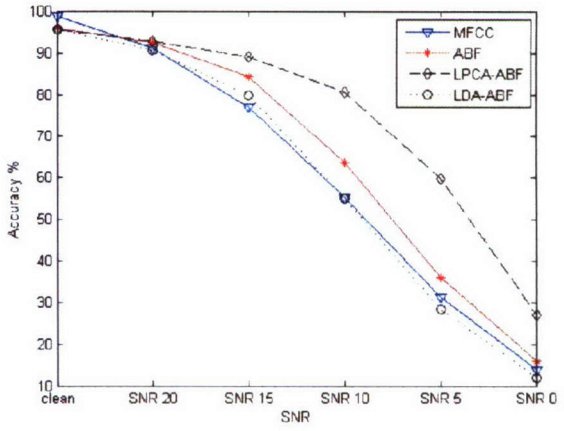

Figure 4. Comparison of ASR performance different feature sets with subway noise.

One of the issues we discussed in the most recent ONR meeting was discrimination ability of features. Thus, we also implemented linear discriminant analysis (LDA) to improve the discrimination between classes, and to reduce dimension of auditory peripheral model output. Here, each digit is associated to a class, so there are $d = 12$ classes i.e., *one* through *nine, zero, oh*, and *silence*. The training data (clean speech) is automatically segmented in word level by using force alignment using acoustic models trained with MFCC since it is more accurate for clean speech. The LDA transformation matrix is learnt from the labeled training corpus, and the output of early auditory system is projected to $(d-1)$ dimensional LDA space using this transformation matrix. The ASR experiments using these features, LDA_APF, are presented in Fig 2 together with previous results for comparison purpose. It can be observed from the figure that using LDA didn't provide any benefit for ASR with clean data. Also, it performed worse comparing to PCA extracted features and ABF when the data used in ASR was noisy. These experiments clearly showed that using PCA in feature extraction for noisy speech recognition is more beneficial comparing to the LDA method.

Another problem we have been working on is creating saliency map and surprise detector for audio and speech signals. We have developed an initial model of auditory saliency model by adapting the models for image and video developed by Itti & Baldi [2005] to audio signals. In this model, first the spectrum of sound is estimated using the early auditory model as previously discussed, and fed to the system as two dimensional time-frequency map. In the second stage this image map is analyzed by extracting a set of features that is similar to those stages in the auditory system. The selected features are intensity, frequency features, temporal features, orientations, and pitch distribution. These features are extracted using spectro-temporal receptive filters mimicking the analysis stages in primary auditory cortex: the intensity filter corresponds to receptive fields with only an excitatory phase selective for a particular region, the frequency feature filters corresponds to receptive fields with an excitatory phase and simultaneous inhibitory side bands, temporal filters corresponds to receptive fields with an inhibitory phase and a subsequent excitatory phase, orientation filters corresponds to neurons sensitive to motion energy, and pitch distribution is calculated from intensity feature map. Multiple scales are created using these spectro-temporal filters, each being a resampled version of the previous.

Then, each feature map is computed by center-surround operation akin to local cortical inhibition. It is implemented in the model by comparing fine and coarse scales. In the next stage, obtained feature maps are normalized and combined across-scales. At the last stage final maps are summed to output the final auditory saliency map. This is the description of our pre-mature model we developed for auditory saliency map. We are presently working on it to tune the parameters, and establish the details of all the parts, and to create appropriate test samples for it.

Summary and Future Work:

- We developed a new set of features by introducing nonlinear compression effect due to the outer hair cells into our auditory model, and then applied PCA. This new set of features provided significant WER improvement in ASR experiments for noisy speech.
- We replaced PCA step in our model with LDA to focus on feature discrimination at the output of our auditory model. The experiment showed that LDA didn't provide any benefit for ASR with clean data, and it was more sensitive to noise comparing to PCA.
- We developed an initial model for auditory saliency map. We are working on it to get a complete working model, and also to find appropriate test sets.
- Our robust ASR research results will be summarized in a paper:
  "Bio-inspired signal processing for robust automatic speech recognition", by *Ozlem Kalinli, Shrikanth Narayanan* (under preparation)
- Our earlier work noise discrimination is accepted for ICASSP 2007:
  "Discriminating two types of noise sources using Cortical Representation and dimension reduction technique", by *Shiva Sundaram, Shrikanth Narayanan*
  Accepted ICASSP 2007. Honolulu, Hawaii, USA.
- Also, our work on the signal representation framework has been accepted at ICASSP:
  Jorge Silva, Shrikanth Narayanan. Optimal Wavelet Packets Decomposition Based On A Rate-Distortion Optimality Criterion. In Proc. ICASSP, Honolulu, Hawaii, April 2007.

4. Developing an integrated perceptual system

Progress with large-scale system integration efforts has been made in two directions. First, we have achieved a near complete specification of the various modules and data structures which will be used in a model that integrates attention, object recognition, rapid computation of the hollistic "gist" of a scene, and a symbolic reasoning back-end. A key new concept which has emerged from this specification work is that of the "semantic representation" of a scene, which is an intermediary between volatile visual representations which change as the scene does, and long-term knowledge. The semantic representation is constructed incrementally as attention scans a scene (with possible biases towards some scene locations derived from gist) and the successively attended locations are identified.

As this sequential process builds up some incomplete and preliminary understanding of the scene, the cognitive back-end in turn biases attention towards desired new targets (e.g., if a hand is attended to, look for the associated face). To achieve this, we have implemented a new variation on the saliency map theme, where a set of desired object features can be specified at the

top (e.g., look for small red elongated objects), and that specification is used at the early visual processing stages to compute a probability that such an object is present at a given image location. We have successfully started applying this approach to simple problems, such as clearing up a table using a robot arm. In this toy problem, a video camera mounted on the wrist of our robot arm first gets a full bird's eye view of the entire workspace. Attention then selects one object, which happens to be the most salient. Low-level visual features of that object are then used to bias the saliency computation while our arm's gripper (and the camera) moves towards the selected object, making the visual servoing of the gripper to the object very simple and robust because the biased saliency map responds much more strongly to the selected object than to other objects or background clutter.

Our system is in its infancy but operates robustly in real time, and is able to clear a desk from all objects placed on it in a reliable manner.

References

Balboa, R.M., and N.M. Grzywacz (2003) Power Spectra and Distribution of Contrasts of Natural Images from Different Habitats. *Vision Res.* **43,** 2527-2537.

Grzywacz, N.M., and J. de Juan (2003) Sensory Adaptation as Kalman Filtering: Theory and Illustration with Contrast Adaptation. Network: Comput. Neural Syst. **14,** 465-482.